Logistic Regression on the BIOSCAD data set

Introduction

**Overview:**

The aim of the project was to create a model that, based on biomarkers measured on the skin, would predict whether a patient suffers from atopic dermatitis (AD). We used the data from the AMC BIOSCAD study.

Method

**Data set:**

The AMC BIOSCAD study data is first pre-processed to make it suitable for training – see pre-processing report (../Preprocessing/preprocessing-report.pdf) for details. One additional step was performed to the pre-processed data – the SCORAD values were replaced with a Boolean to indicate if the patient suffered from AD. After pre-processing, there were 35 attributes that could be used for training, 1 output attribute, and 100 data points.

**Model building:**

Building each model consisted of a training, validation, and testing phase. To train the model, 60% of the data was used. Using the validation data (20%), the best prediction threshold was found. This is the value above which a patient is predicted to have AD, and below which not. Once the best prediction threshold has been found, the data is tested using the remaining 20% to allow the performance of the model to be evaluated. This 3 step process is repeated 100 times using different data points for the training, testing, and validation data.

**Performance evaluation:**

Several performance metrics were calculated to test the models. The first and most used of which is the accuracy – the percentage of predictions that were correct. We also calculated the sensitivity (number of true positives divided by actual positives), and the specificity (number of true negative actual negatives). We also computed the F-score, the Mathews correlation coefficient and the diagnostic odds ratio.

**Input variable selection:**

The first model built used all 35 attributes as inputs. To see the relative importance of these attributes, we also built models using each attribute individually. Using this information, we combined the top *n* performing variables (based on accuracy) together to create input data for new models. These showed us that, in general, the greater the number of top performing attributes used, the worse the performance of the model (see results section). Because of this, we then built every possible combination of models using up to 5 attributes as inputs. This is approximately 400,000 different models.

We also performed logistic regression using only IL-1a, IL1β as these were the only two attributes in which most of data was above the detection limit.

Results

**Top performing models:**

<Table with 15 best models>

**Most accurate model:**

<Table with details for most important values>

**Reliable attributes model:**

Discussion

**Input attributes:**

The most accurate model used 3 attributes as inputs. These were FLG-Carrier, IL-7, and IL-2. The 10 next most accurate models also used these 3 inputs and one additional input. This indicates that these 3 attributes have the largest effect on whether or not a patient is predicted to have AD.

**Coefficient analysis:**

The coefficients of the best model (table x) show that FLG-Carrier had the most significant impact on the result of the prediction. This is because in the original data set, there are no patients with FLG mutations in the control group. As such, the model indicates that any patient with an FLG mutation must have AD. However, we know from other analysis of other data sets that this is not the case. Thus, the model is overfitted to the input data and will not generalise well to external data.

**Input data quality:**

Another issue is that the majority of data in the BIOSCAD data set is marked as below the detection range. As such, the data is unreliable and thus the models also unreliable (garbage in, garbage out principle). This is shown as the best selected continuous attributes (IL-2, and IL-7) have already been shown to have limited impact on SCORAD.